Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning

Ahmed Salem, Apratim Bhattacharya, Michael Backes
Mario Fritz, Yang Zhang

CISPA Helmholtz Center for Information Security, Max Planck Institute for Informatics
Online Learning

- Data generation rate
- 90% of the data in the world today has been created in the last two years alone
- Cost of retraining
Attack Surface in Online Learning

Research Question:
Can this posterior difference be a new attack surface?
Threat Model

- Attacker has black-box access to the target model
- Attacker knows:
  - Target model’s architecture
  - A shadow dataset from the same distribution of the target model’s dataset
General Attack Pipeline

- **Probing set**
  - 7

- **Target Model**

- **Update**
  - Probing set
  - 7

- **Posterior difference**

- **Attack Model**
  - Single-sample label Inference
  - Single-sample reconstruction
  - Multi-sample label distribution
  - Multi-sample reconstruction
Attack Model Training

- Target model’s architecture
- Shadow dataset

Target Model

Probing Set

Shadow Model

Shadow Updated Model 1

Shadow Updated Model n

Posterior difference 1

Posterior difference n

X

Y

updating set 1

updating set 2

updating set n
Single-sample Label Inference

It is a 0

• More than 6x and 9x better than baseline for MNIST and CIFAR-10
Single-sample Reconstruction

- More complicated than inferring label
- Attacker needs a sample generator
  - We rely on autoencoder's decoder
Autoencoder
Single-sample Reconstruction

\[ A_{SSR} \]

\[ \delta \rightarrow \text{Encoder} \rightarrow \mu \rightarrow \text{Decoder} \rightarrow x_{update} \]

Autoencoder

\[ D_{shadow} \rightarrow \text{Encoder} \rightarrow \text{Decoder} \rightarrow D_{shadow} \]

Transfer
Single-sample Reconstruction

Mean squared error (MSE)

Autoencoder (Oracle)

CIFAR-10

MNIST
Multi-sample Label Estimation

Probing set

Target Model

Update

Probing set

Attack Model

KL-divergence as the loss

Multi-sample label distribution

KL-divergence as the loss

Probing set

Update

Target Model

KL-divergence as the loss

Multi-sample label distribution
Multi-sample Label Estimation

KL-divergence

$A_{LDE}$
Baseline
Transfer 10-100

MNIST (10)
CIFAR-10 (10)

MNIST (100)
CIFAR-10 (100)
Multi-sample Reconstruction

- Most challenging scenario in this attack scenario
- Reconstruct a set of data samples
  - Autoencoder cannot help anymore
- What we do?
Generative Adversarial Network (GAN)

Image credit: Thalles Silva
Multi-sample Reconstruction

\[ \mathcal{L}_{BM} = \sum_{\hat{x} \sim G} \min_{x \in D_{\text{update}}} \|\hat{x} - x\|_2^2 + \sum_{\hat{x}} \log(D(\hat{x})) \]
Multi-sample Reconstruction

Mean squared error (MSE)

- One-to-one match
- $A_{MSR}$
- Baseline

MNIST

CIFAR-10
Multi-sample Reconstruction
Multi-sample Reconstruction
Summary

Probing set

Target Model

Update

Probing set

Attack Model

Encoder

Decoder

Posterior difference

Thank you for your attention!
Questions?

ahmed.salem@cispa.saarland
https://ahmedsalem2.github.io/
@AhmedGaSalem